# Do Flexible Administration Procedures Promote Individualized Clinical Assessments? An Explorative Analysis of How Clinicians Utilize the Funnel Structure of the SCID-5-AMPD Module I: LPFS

Aleksander Heltne, Johan Braeken, Benjamin Hummelen, Sara Germans Selvik, Tore Buer Christensen & Muirne C. S. Paap

**Routledge**
Taylor & Francis Group

# Do Flexible Administration Procedures Promote Individualized Clinical Assessments? An Explorative Analysis of How Clinicians Utilize the Funnel Structure of the SCID-5-AMPD Module I: LPFS

Aleksander Heltne[1,2] (iD), Johan Braeken[1,3] (iD), Benjamin Hummelen[1] (iD), Sara Germans Selvik[4,5] (iD), Tore Buer Christensen[6] (iD), and Muirne C. S. Paap[1,7] (iD)

[1]Department of Research and Innovation, Clinic for Mental Health and Addiction, Oslo University Hospital, Oslo, Norway; [2]Institute of Clinical Medicine, University of Oslo, Oslo, Norway; [3]Centre for Educational Measurement, University of Oslo (CEMO), Oslo, Norway; [4]Department of Psychiatry, Helse Nord-Trøndelag, Namsos Hospital, Namsos, Norway; [5]Department of Mental Health, Norwegian University of Science and Technology (NTNU), Trondheim, Norway; [6]Department of Mental Health, Sørlandet Hospital, Arendal, Norway; [7]Department of Child and Family Welfare, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands

## ABSTRACT

The current study examined clinicians' utilization of the SCID-5-AMPD-I funnel structure. Across 237 interviews, conducted as part of the NorAMP study, we found that clinicians administered on average 2-3 adjacent levels under each subdomain, effectively administering only about 50% of available items. Comparing administration patterns of interviews, no two interviews contained the exact same set of administered items. On average, when comparing individual interviews, only about half of the administered items in each interview were administered in both interviews. Cross-classified mixed effects models were estimated to examine the factors affecting item administration. Results indicated that the interplay between patient preliminary scores and item level had a substantial impact on item administration, suggesting clinicians tend to administer items corresponding to expected patient severity. Overall, our findings suggest clinicians utilize the SCID-5-AMPD-I funnel structure to conduct efficient and individually tailored assessments informed by relevant patient characteristics. Adopting similar non-fixed administration procedures for other interviews could potentially provide similar benefits compared to traditional fixed-form administration procedures. The current study can serve as a template for verifying and evaluating future adoptions of non-fixed administration procedures in other interviews.

Semi-structured clinical interviews such as the Structured Clinical Interview for DSM-5 (SCID-5-PD) and the Structured Clinical Interview for the DSM-5 Alternative Model for Personality Disorders (SCID-5-AMPD) (First et al., 2018b, 2015a) have long been the gold standard for assessment of personality disorders in both research and clinical settings. Such interviews have been shown to be more reliable than unstructured clinical interviews, and to be better at differentiating patients from non-patients than self-report measures (Ekselius et al., 1994; Miller et al., 2001; Ottosson et al., 1998; Saigh, 1992). An important advantage of semi-structured clinical interviews is their ability to strike a balance between the standardization and personalization of a clinical assessment (Mueller & Segal, 2015; Widiger & Samuel, 2005). By providing a structured framework for what an assessment should cover, semi-structured interviews ensure a consistency in information gathering and a comparability of scores which unstructured clinical interviews cannot match. Similarly, by allowing clinicians to follow up

on and evaluate patient responses, they provide a more personalized and in-depth assessment than self-report measures or structured interviews (e.g., clinician-rated questionnaires). Furthermore, semi-structured interviews also reduce the risk of over- and underreporting of symptoms (Ottosson et al., 1998; Stuart et al., 2014).

The legacy instruments SCID-II and SCID-5-PD follow fixed administration rules in which all diagnostic criteria for all PD diagnostic categories are assessed, and interview items (i.e., interview questions) are administered in a chronological order (First & Gibbon, 2004; First et al., 1997, 2015a, 2015b). The SCID-5-AMPD Module I Level of Personality Functioning Scale, developed to assess Criterion A of the DSM-5 Section III Alternative Model for Personality Disorders (AMPD), on the other hand, follows a so-called funnel structure, whereby initial screening questions are used to inform decisions about which parts of the interview to administer, giving clinicians more freedom to choose which items to administer, as well as the order in which to

administer them (First et al., 2018a). While this freedom provides an opportunity to conduct more individually tailored assessments than what would be possible with fixed administration procedures, it is important to explore whether clinicians make use of that flexibility, and whether their selection of interview content is based on clinically relevant factors.

The administration of all items in the SCID-II and SCID-5-PD is useful for multiple reasons. Firstly, as the standard categorical model on which these interviews are based allow for the concurrent diagnosis of multiple personality disorders (American Psychiatric Association, 1994, 2013); and since the presence of multiple concurrent personality disorders is quite common in clinical practice (Clark, 2007), the assessment of all diagnostic categories is often appropriate. Secondly, the diagnosis of Personality Disorder Not Otherwise Specified (PD-NOS), is commonly used if patients satisfy general criteria for personality disorder and multiple specific PD symptom criteria, without qualifying for a specific PD diagnosis. As such, an overview of the types of personality disorder criteria a patient qualifies for can be valuable diagnostic information. In the AMPD, however, personality disorder diagnoses are redefined in terms of personality functioning and pathological personality traits (American Psychiatric Association, 2013). Personality functioning is operationalized as the levels of personality functioning (LPFS). The LPFS describes self- and interpersonal functioning across the four domains of Identity, Self-Direction, Empathy, and Intimacy. Five levels of functioning are outlined under each of these domains ranging from 0 (little to no impairment) to 4 (extreme impairment). Each level is defined by a description of what functioning on that level under a given domain entails. Assessing the LPFS involves evaluating the appropriateness of the descriptions associated with the five levels of functioning under each of the four domains. Only one level of functioning can be assigned to each domain. This means that, rather than assessing all levels for each domain, the clinician only needs to identify the one level that best fits the patient.

The SCID-5-AMPD Module I: LPFS (SCID-5-AMPD-I) is the SCID-II and SCID-5-PD equivalent for the LPFS (First et al., 2018a). Highlighting descriptive elements outlined in the original LPFS, this interview divides the four domains of the LPFS (Identity, Self-direction, Empathy and Intimacy) into 12 subdomains (three for each of the four domains). The interview begins with the clinician posing eight general questions about the patient; how they view themselves and their relationships. Based on the answers to these eight initial questions, the clinician assigns a preliminary score between 0 and 4, representing the five levels of personality functioning outlined in the LPFS. The clinician then moves on to assess individual subdomains. When assessing a subdomain, screener questions are used, alongside the preliminary score assigned at the beginning of the interview, in order to estimate which level of functioning may best describe the patient under a given subdomain. Based on this initial evaluation, the clinician selects a level, and proceeds to pose interview questions corresponding to this level under the subdomain being assessed (these interview questions listed under each level of each subdomain will henceforth be referred to as *level-specific items*). Level-specific items are used to further evaluate which level may best apply to the patient, and should be administered unless a) the clinician has already gathered sufficient information in order to answer them; or b) it becomes apparent that the level under assessment is not appropriate to describe the patient, suggesting the clinician should move to another level on the LPFS scale. In this way, the clinician is instructed to move between levels (e.g., from 1 to 2 or from 2 to 4), when the information gathered during the interview suggests that it is appropriate to do so. In order to ensure that the clinician does not underestimate the severity of the patient's condition, the clinician is also instructed to continue to assess increasing levels of impairment until the patient clearly no longer qualifies for the level being assessed. The assessment of subdomains continues in this manner until the clinician decides which level description is best suited to describe the patient's functioning in each subdomain (First et al., 2018a). This means that clinicians are not obliged to administer interview questions to which they already have obtained the answers, or which they consider not to be relevant or applicable to the patient.

This funnel structure provides flexibility by allowing the clinician to select which parts of the interview to administer to a given patient. For instance, if a certain level under a given subdomain clearly is not suited to describe the patient, then this level may not need to be administered. This flexibility has the potential to make the interview more efficient and individualized than a standard, structured interview with fixed response categories, in which all items are administered. In the SCID-5-AMPD-I, clinicians are only required to administer the parts of the interview that they consider relevant to a specific patient. It is clear, that the flexible approach (whereby items that are not deemed relevant can be discarded) clashes with the standard approach (whereby all items are administered sequentially in a predefined order). At the time of writing, the implications of introducing this novel administration procedure have only been partially examined. Previous research has primarily focused on the psychometric properties of the SCID-5-AMPD-I, documenting good to excellent interrater reliability in clinical samples at the level of subdomain, domain and total scores (Buer Christensen et al., 2018; Fossati & Somma, 2021; Meisner et al., 2022). The instrument has also been found by Hummelen et al. (2021) to be unidimensional, supporting the interpretation of the total score as a single construct. Furthermore, the SCID-5-AMPD-I has been shown to be suitable for differentiating PD patients from healthy controls (Buer Christensen et al., 2019; Meisner et al., 2022), and to be a good predictor of outcome measures related to PD severity, such as number of previous suicide attempts (Kampe et al., 2018).

Although these findings are indeed promising, it is still unclear whether the funnel structure makes the SCID-5-AMPD more efficient and individually tailored than an interview in which all items need to be administered.

Furthermore, there are no studies examining the factors affecting clinicians´ decisions about which sections of the interview to administer (e.g., which levels clinicians choose to assess, and which level-specific items they choose to administer). The funnel structure of the SCID-5-AMPD-I may pose challenges to the clinicians regarding their ability to make sound clinical decisions (based on relevant patient characteristics) during the administration of the interview. For instance, the structure of the interview itself or the clinicians´ preferences for certain kinds of information could unduly influence the decisions. Moreover, clinicians may avoid administering confrontational interview items or be more inclined to administer sections of the interview for which there are many listed items, as more items could be taken to imply that a section has greater importance. Since clinicians´ decisions about which parts of the interview to administer are likely to impact scoring and subsequent diagnostic conclusions, we believe it is important to examine the factors influencing these decisions. If the funnel structure is found to function as intended (i.e., promoting efficiency and individualized clinical assessments by allowing clinicians to tailor the interview to individual patients), and the selection of interview content is found to be determined by relevant patient characteristics, this would suggest that similar funnel-structured administration procedures could potentially be used in other interviews where the underlying diagnostic criteria are suitable (e.g., where the subject of assessment is continuous). The current study could then serve as a template for evaluating whether such adoptions are successful.

In the current study, we seek to explore whether clinicians make use of the possibility offered by the funnel structure of the SCID-5-AMPD-I to shorten the assessment and to tailor it to individual patients, and to examine which factors contribute to the selection of interview items. To this end, we will examine data from SCID-5-AMPD-I interviews administered by clinicians in The Norwegian Study of the AMP (NorAMP), which, at the time of writing, is the largest SCID-5-AMPD-I dataset on record. We aim to answer the following research questions:

1. To what extent do clinicians limit the SCID-5-AMPD-I assessment to only the levels of functioning under each subdomain considered to be appropriate for the patient being assessed? To answer this research question, we will examine the average number of levels assessed under each subdomain of the interview. If clinicians follow the administration procedures of the interview, we do not expect them to administer level-specific items from all levels under each subdomain. Rather, we expect that clinicians primarily administer level-specific items from a subset of adjacent levels.
2. To what extent do clinicians tailor the administration of interview content to the patient being assessed? To answer this research question, we will attempt to identify patterns of item administration across the dataset. Since we assume that clinicians are actively adapting the administration of interview content to the patient being

**Table 1.** Number of participants per personality disorder.

| Personality disorder | N |
| --- | --- |
| Avoidant | 57 |
| Dependent | 9 |
| Obsessive-compulsive | 16 |
| Paranoid | 22 |
| Schizotypal | 0 |
| Schizoid | 2 |
| Histrionic | 1 |
| Narcissistic | 2 |
| Antisocial | 20 |
| Borderline | 52 |

assessed, we expect to find a multitude of unique administration patterns.

3. What factors contribute to the selection of interview items? To answer this research question, we will estimate a cross-classified generalized mixed effects model (which in many ways resembles an item response theory (IRT) model) exploring the effects of item characteristic and expected patient severity on item administration. Here, we expect item administration to be impacted by an interaction between clinicians' expectations about the severity of their patients' impairment in personality functioning (as indicated by preliminary scores assigned at the beginning of the interview) and the associated severity of individual items (as indicated by the level under which an item is listed in the interview). Essentially, we expect clinicians to administer items that correspond to the preliminary scores assigned to the patient being assessed, as instructed by the interview guidelines. Other item characteristics (item length, the order in which items appear in the interview, and the number of alternative items available) on the other hand, are not expected to play a major role in item administration, as these represent more superficial characteristics of the interview content and structure.

## Methods

### Participants and procedure

The data for this study were collected as part of the Norwegian Study of the Alternative Model for Personality Disorders (NorAMP). The NorAMP study includes 317 participants, 282 psychiatric patients and 35 healthy controls. Participants in the clinical sample were recruited from different levels of psychiatric care across four Norwegian hospitals. Participants in the healthy control sample were recruited through information posters at the University of Agder, the University of Oslo, and Sorlandet Sykehus. During the data collection for the NorAMP study, patients were administered the SCID-5-AMPD-I. For the purposes of the current study, we extracted data on which items had been administered in 237 of these interviews. Our sample consisted of 152 women and 85 men, and the mean age in the sample was 31.63 ($SD = 10.12$). Most patients ($n = 142$) qualified for at least one personality disorder according to the SCID-II at the time when the SCID-5-AMPD-I interviews were conducted (range $= 0—7$, median $= 1$). The most common personality disorders in the sample were

Avoidant Personality Disorder ($n = 57$) and Borderline Personality Disorder ($n = 52$). See Table 1 for a full overview of all DSM-IV personality disorder diagnoses and their prevalence in the current sample. SCID-5-AMPD-I interviews were carried out by twelve licensed psychologists and psychiatrists from four Norwegian hospitals, who contributed to the data collection of the NorAMP study. The twelve licensed psychologists and psychiatrists had all received a two-day training seminar covering the theoretical background of the DSM-5-AMPD and the LPFS as well as case-based training sessions for the administration and scoring of the SCID-5-AMPD-I. In addition to this two-day training seminar, raters held regular calibration meetings throughout the data collection period for the NorAMP study.

## Ethics

The data for this study were collected as part of the NorAMP study. The NorAMP study was approved by the regional committee for medical and health research ethics (REK) on January 12, 2015 (REK ref.no: 2014/1696). An official notice was submitted to REK regarding the additional analyses conducted for the purposes of this study and was approved on January 20, 2022 (REK ref.no: 9808).

## Measures

All variables used in this study were extracted from data from the NorAMP study. We extracted data on who had conducted the interview (Interviewer), clinicians initial expectations of patients' personality impairment (Preliminary LPFS scores), characteristics of the items being administered (Item characteristics), and the administration of interview content (Item administration, Level administration and Administration patterns).

## Interviewers

In the dataset used for the current study, each clinician had been assigned a code ranging from 1 to 12 to ensure anonymity.

## Preliminary LPFS score

The SCID-5-AMPD-I is a semi-structured clinical interview which assesses the Levels of Personality Functioning scale which make up the criterion A of the DSM-5-AMPD. The interview consists of 8 initial questions, 28 screening questions (between 1 and 5 per subdomain), and 168 level-specific items - i.e., interview questions listed under each level under each subdomain (between 1 and 6 listed under each level under each subdomain). The interview follows an administration procedure whereby the clinician first administers the 8 initial questions. Based on these initial questions, the clinician assigns the patient a preliminary LPFS score between 0 (indicating little or no impairment) and 4 (indicating extreme impairment), corresponding to the 5 levels in the LPFS. This preliminary LPFS score and the screening questions listed under each sub-domain are used as a guide to determine which levels to assess and which of the 168 level-specific items to administer throughout the interview. Preliminary LPFS scores, ranging from 0 to 4, assigned to patients at the beginning of the SCID-5-AMPD-I interview were recorded as an indication of clinicians' expectations about the patients' level of personality impairment.

## Item characteristics

For all 168 level-specific items in the SCID-5-AMPD-I, available item characteristics were extracted from the Norwegian translation of the SCID-5-AMPD-I interview used in the NorAMP study. The variable *Item length* denotes the length in characters (ranging from 27 to 229) of each item. The variable *Item level* denotes the level (ranging from 0 to 4) under which an item is listed under a given subdomain. The variable *Item alternatives* denotes the number of items available (ranging from 1 to 6) under a given level of a given subdomain. Lastly, the variable *Item order* denotes the order (ranging from 1 to 6) in which an item appeared on a list of available items under a given level of a given subdomain.

## Item administration

In the NorAMP study, prior to administering the SCID-5-AMPD-I, clinicians were instructed to record patient responses to all administered interview items, assigning a number to signify whether the patient confirmed (3), partly confirmed (2) or denied (1) the content of an administered item. For the purposes of the current study, we were interested in whether or not individual interview items had been administered. Recorded patient responses were therefore re-coded into a dichotomous variable, item administration, where $0 =$ not administered and $1 =$ administered. Item administration was recorded for all 168 level-specific items across 237 available SCID-5-AMPD-I interview protocols where recordings of patient responses were consistent and legible. A trained graduate student, under supervision of the first author, looked through all available pen-and-paper interview protocols from the NorAMP study. Any item for which a response marking could be located was coded as administered, and any item for which a response marking could not be located was coded as not administered. Items for which response markings were illegible or ambiguous (e.g., arrows or markings drawn next to an item without any recorded patient response) were coded as missing, as it was impossible to determine whether the item had been administered or not.

In addition to having access to pen-and-paper interview protocols for the interviews included, we also had access to twelve video-recorded interviews. In order to evaluate whether the indirect scoring of item administration could be considered accurate, observational scoring of item administration was conducted for the twelve video-recorded interviews available to us. A trained graduate student, under supervision of the first author, watched all twelve interviews in their entirety and scored item administration observationally. Though the twelve interviews were conducted by only four out of the twelve raters involved in the NorAMP study,

these four raters had conducted the majority (55.27%) of the interviews included in the current dataset. Across these interviews, we found between 1 and 17 items for which interviewers had recorded a response marking in their pen-and-paper interview protocol for an item which was not observed as being administered in the video-recorded interview (Median = 5).

Similarly, we found between 0 and 16 items for which interviewers had not recorded a response marking in their pen-and-paper interview protocol for an item which was observed as being administered in the video recorded interview (Median = 2). Both kinds of errors in clinicians´ reporting may have impacted the overall quality of our dataset, and lacking access to video-recordings of all included interviews, it is difficult to determine the overall rate of erroneous reporting. Given that no clear pattern was observed for these errors in reporting, and because the incidence rate of erroneous reporting was low across these twelve interviews, the item administration data were assumed to be of acceptable quality.

### Level administration

For each level under each subdomain of each interview, level administration was recorded as a dichotomous variable (0 = not administered, 1 = administered). Across interviews, a level under a given subdomain was recorded as administered if at least one level-specific item from that level had been administered under that subdomain.

### Administration patterns

For each interview, administration patterns were recorded as a 168-digit numeric pattern representing item administration across all 168 level-specific items. For an interview in which the first three level-specific items had all been administered, the first three digits in the administration pattern of that interview would be '111'. Conversely, if the first three level-specific items had not been administered in an interview, the first three digits in the administration patterns of that interview would be '000'.

### Analysis

All analyses were performed in R version 4.1.2 (R Core Team, 2021). In order to evaluate whether clinicians in our sample shortened the SCID-5-AMPD-I interview by administering items from only a subset of adjacent levels for each subdomain (RQ1), we examined the number of levels administered under each subdomain of each interview, as well as the number of adjacent levels administered per subdomain across interviews.

In order to evaluate whether clinicians in our sample adapted the administration the SCID-5-AMPD-I interview to individual patients (RQ2), administration patterns were compared across interviews. The amount of overlap between interviews was calculated as median percent overlap between each administration pattern against all other administration patterns. Essentially, we compared the administration

pattern of each individual interview to that of every other interview (236 comparisons per interview), calculating the pairwise percent overlap in administration patterns. The median of these 236 percentages, for each interview is labeled the median percent overlap for that interview. A Monte Carlo simulation test (Mooney, 1997) was performed to compare the amount of observed overlap in administration patterns to the expected overlap if item administration had been random (50% likelihood of administration for each item).

In order to examine the factors that contributed to the selection of items for administration during the interview (RQ3), we modeled the probability of administering an item as a function of patient, item, and interviewer characteristics using a mixed-effects logistic regression modeling approach. A model-comparison strategy using likelihood-ratio tests was implemented and used to compare a sequence of four models which incorporated a) study design factors, b) interview-procedural factors, and c) more incidental item features. Models were estimated using full-information maximum likelihood through the lme4 R-package (Bates et al., 2015). Relative variance components are reported as $R^2$ effect size measures for random effects and joint fixed effects separately, and for the full model.

In a first baseline model, we accounted for the fact that interviewers might vary on the number of questions they tend to ask, that patients might vary on the number of questions they would prompt the interviewers to ask and elicit, and that items might vary in the tendency to be askedi. A variance component (i.e., random intercept) was included for each of these three main sources of variation in item administration that are a natural part of the study design. In a second model, a fourth interviewer-by-item variance component was added to reflect the possibility that interviewers might differ in their tendency to administer specific items. Given the training all twelve interviewers had received, we did not expect large differences in their personal implementation of the interviewing procedures; however, it is still not unlikely that some idiosyncratic tendencies could occur.

In order to assess whether the intended individualization of the interview occurred, matching the level of the administered item to initial expectations of the patient's severity as the interview guidelines prescribe, the interaction between the level of the item and the SCID-5-AMPD-I preliminary score were added as fixed effects to a third model. In order to study the relative impact of the three more incidental item features over and above the more structural interview-procedural factors, the three item characteristics (order, alternatives, length) were added as fixed effects to a fourth model.

## Results

### Missing data

Across interviews, the rate of missing data for the item administration variable ranged from 0-17%, with a median of 1%. Only 6 out of 237 interviews had more than 10% missing values on the item administration variable. Across items, the rate of missing data ranged from 0-16% with a median of 1%, with only 2 out of 168 items having more

than 10% missing. Examining the 10 items with the highest rate of missing values, we found that 8 of these items covered narcissistic or antisocial behaviors or qualities (e.g., 'Do you think other people should just get out of your way, and let you do what you want?' or 'Are you willing to ignore the rules to get what you want?'). This may indicate that clinicians were reluctant to administer confrontational items or that clinicians deemed these items to be non-relevant. Beyond these observations no systematic tendencies were found for the missing data. Missing values on the item administration variable were treated as missing at random.

For the Preliminary LPFS score variable, 25 interviews had missing values. These 25 interviews were therefore excluded when estimating the cross classified mixed effects model. For the item characteristic variables (item length, item level, item order, and item alternatives), there were no missing values.

## Descriptive statistics

Across the 237 interviews, between 22 and 165 out of 168 level-specific items had been administered. The mean number of administered items was 84, which constitutes 50.1% of the 168 available level-specific interview questions. On average, individual items were administered in 119 interviews out of 237 included interviews (range: 16 - 187). Table 2 shows the median and range for all predictor variables included in the cross classified mixed effects model.

## Level administration and level adjacency

Clinicians administered items from between 0 and 5 levels of severity under each subdomain. Table 3 shows an overview of mean number of levels administered under each subdomain. This table also shows, for each subdomain, the number of interviews in which clinicians administered 2, 3, 4 and 5 adjacent levels. Lastly, the table shows the number of interviews in which clinicians administered items from non-adjacent levels or administered no level-specific items. As can be seen from this table, the mean number of levels administered ranged from 2.34 to 3.07 across the 12 subdomains. Across the 12 subdomains, there were between 0 and 4 interviews in which clinicians administered no level-specific items. This was surprising as it would indicate that for these interviews, patients were assigned scores for a given subdomain without the clinicians actively assessing any of the five levels of severity. This occurred most frequently for subdomain 9 (Understanding of Effects of Own Behavior on Others). Similarly, across subdomains there were between 4 and 25 interviews in which clinicians administered level-specific items from all five levels. This would suggest that for these interviews, clinicians may have had trouble determining a patient's score on a given subdomain, and therefore found it necessary to actively assess all levels of severity. This occurred most commonly for subdomain 11 (Desire and Capacity for Closeness). Instances of both none and all levels being administered were very rare.

By far the most common observation under each subdomain was that clinicians administered items from 2 or 3 adjacent levels. This would suggest clinicians carried out an active evaluation of a subset of adjacent levels before determining a final score for the patient, as is described in the administration guidelines for the interview. There were however between 3 and 26 interviews per subdomain for which clinicians administered items from two or more non-adjacent levels. In these instances, clinicians most commonly administered items from levels 0 and 2, skipping level 1, or administered items from levels 1 and 3, skipping level 2. These two patterns were observed for 111 out of 118 instances of non-adjacent item administration. Considering there are in total 2844 level administration instances (12 subdomains by 237 participants), non-adjacent level administration was also very rare. Thus, in most cases clinicians administered items from

**Table 2.** Median and range for predictive variables.

| Classification | Variable name | Median | Range |
|---|---|---|---|
| Item characteristics | Item level | 2 | 0–4 |
| | Item order | 2 | 1–6 |
| | Item alternatives | 3 | 1–6 |
| | Item length | 84 | 27–229 |
| Patient severity | Preliminary score | 2 | 0–4 |

**Table 3.** Overview of level assessment across subdomains.

| Subdomains | Mean | 2 Adj | 3 Adj | 4 Adj | 5 Adj | Nonadjacent admin | Non admin |
|---|---|---|---|---|---|---|---|
| ● Identity | | | | | | | |
| Sense of Self | 3.00 | 55 | 90 | 59 | 18 | 10 | 0 |
| Self Esteem | 3.07 | 50 | 97 | 71 | 9 | 3 | 0 |
| Emotional range and regulation | 2.81 | 78 | 86 | 52 | 7 | 6 | 0 |
| ● Self-direction | | | | | | | |
| Ability to pursue meaningful goals | 2.58 | 83 | 81 | 33 | 10 | 9 | 1 |
| Internal standards of behavior* | 2.65 | 89 | 80 | 39 | 8 | 8 | 0 |
| Self-reflective functioning | 2.66 | 77 | 83 | 43 | 8 | 9 | 2 |
| ● Empathy | | | | | | | |
| Appreciation others′ experiences** | 2.87 | 69 | 80 | 56 | 14 | 7 | 3 |
| Tolerance of differing perspectives | 2.76 | 72 | 77 | 50 | 13 | 8 | 2 |
| Understanding effect own behavior*** | 2.65 | 72 | 95 | 41 | 4 | 8 | 4 |
| ● Intimacy | | | | | | | |
| Depth and duration of connections | 2.81 | 59 | 80 | 48 | 19 | 12 | 0 |
| Desire and capacity for closeness | 2.73 | 73 | 65 | 40 | 25 | 11 | 3 |
| Mutuality of regard reflected in behavior | 2.34 | 80 | 61 | 37 | 7 | 26 | 1 |

*Note.* Mean = mean number of levels assessed per subdomain, n Adj = number of interviews for which clinicians administered interview items from n adjacent levels. Nonadjacent admin = number of interviews for which clinicians administered items from nonadjacent levels. Non admin = number of interviews for which clinicians did not assess any levels under a given subdomain. The complete labels of these subdomains are: *constructive, prosocial internal standards of behavior; **Understanding and appreciation of others′ experiences and motivations; and ***Understanding of effects of own behavior on others.

between 2 and 3 adjacent levels under each subdomain, and deviations from this pattern were rare.

## Analysis of administration patterns

Across interviews, all administration patterns were unique, with no two interviews having the exact same value for item administration across all items. This was surprising, as it suggests no two patients were administered exactly the same set of interview items. Across interviews, average median percentwise overlap ranged from 6.7 to 97.6 and the average median overlap across all interviews was 55.7%. This observed average median overlap across interviews indicates that, on average, interviews in our sample share about half of the administered content with other interviews. If the SCID-5-AMPD-I interview had followed a fixed administration procedure like the one specified for SCID-II and SCID-5-PD, we would expect to find 100% overlap across interviews. Figure 1 shows the percentwise overlap across all interviews. For each interview, a thin boxplot was drawn of their percentwise overlap with all other interviews. Interviews were organized in ascending order of overlap from left to right, with interviews at the leftmost end of the graph having the least median percentwise overlap with other interviews,
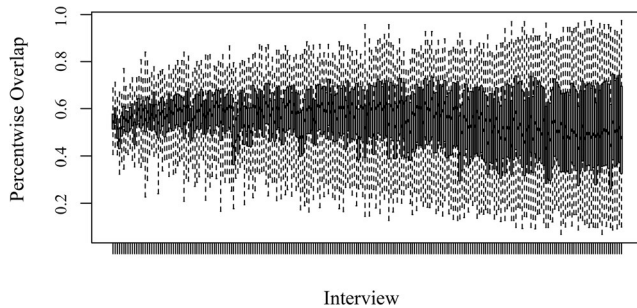


**Figure 1.** Overlap in administraton pattern across interviews.
Note. Each horizontal line represents a boxplot of the percentwise overlap of a single interview against all other interviews. Interviews are organized in ascending order of overlap from left to right.

and interviews at the rightmost side having the greatest median percentwise overlap with other interviews.

Observed overlap was found to have a greater range than expected overlap if item administration had been random, with observed minimum overlap (6.7%) being less than minimum expected random overlap (48.5), and observed maximum overlap (97.6%) being greater than maximum expected random overlap (70.7). This degree of diversity in observed overlap suggests that observed overlap among administration patterns in our sample was systematic and not likely to occur at random, which implies that clinicians adjust the administration of interview content to individual patients.

## Factors influencing item administration

Table 4 outlines the findings of the cross-classified generalized mixed effects analysis. In the initial baseline model (Model 1), the main differences in item administration due to interviewer, patient and item variances were accounted for. About 4% of the variation in item administration could be attributed to differences among interviewers (i.e., some interviewers tend to administer slightly more/fewer items than others), about 11% could be attributed to differences among patients (i.e., some patients may prompt the interviewer to ask and elicit more/fewer questions than other patients), and about 14% to differences between items (i.e., some items tend to be administered more/less frequently than others). The probability of an average item being administered by the average interviewer to the average patient was estimated to be about 50% (i.e., logit($\beta_0 = -.06$)). This lines up with the observation, mentioned under descriptive statistics, that patients were, on average, administered half of the 168 level-specific interview items.

In order to account for the possibility that item administration varied across interviewers, the next modeling step (Model 2), included an item-by-interviewer interaction term, added to reflect the extent to which interviewers differ in their tendency to administer certain items above their personal trend in the overall rate of item administration. This

**Table 4.** Cross-classified generalized mixed effects model for item administration.

| | Model 1 | | Model 2 | | Model 3 | | Model 4 | |
|---|---|---|---|---|---|---|---|---|
| Fixed effect | Coefficient | S.E. | Coefficient | S.E. | Coefficient | S.E. | Coefficient | S.E |
| Intercept | −.06 | .16 | −.05 | .17 | 4.43 | .32 | 4.15 | .36 |
| Item level | | | | | −2.72 | .08 | −2.79 | .08 |
| Preliminary score | | | | | −2.04 | .09 | −2.04 | .09 |
| Item level-by-Preliminary score | | | | | 1.22 | .02 | 1.22 | .02 |
| Item order | | | | | | | −.36 | .09 |
| Item alternatives | | | | | | | .37 | .08 |
| Item length | | | | | | | .20 | .08 |
| Random effect | Variance | RVC | Variance | RVC | Variance | RVC | Variance | RVC |
| Interviewer | .17 | .04 | .20 | .04 | .36 | .04 | .36 | .04 |
| Patient | .49 | .11 | .62 | .11 | .98 | .10 | .98 | .10 |
| Item | .64 | .14 | .63 | .11 | 1.25 | .13 | 1.05 | .11 |
| Item-by-Interviewer | | | .77 | .14 | .59 | .06 | .59 | .06 |
| Fixed effects R² | 0 | | 0 | | .34 | | .36 | |
| Total RVC | .29 | | .40 | | .33 | | .31 | |
| Residual variance | .71 | | .60 | | .33 | | .33 | |
| Log likelihood | −20871 | | −19964 | | −16454 | | −16440 | |
| Number of parameters | 4 | | 5 | | 8 | | 11 | |

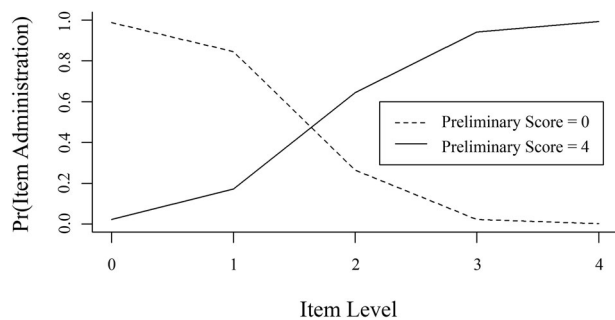*Note.* R² = R-squared, RVC = relative variance component.

**Figure 2.** Probability of administering items from a given level for patients with high and low preliminary scores.
Note. Pr (Item Administration) = The probability of item administration.

item-by-interviewer variance component accounted for about 14% of the variance in item administration. The relative variance component of items dropped from 14% to 11% after introducing the item-by-interviewer variance component to the model while other variance components remained stable. This finding indicates that clinicians in our sample differed in their tendency to administer certain items ($\chi^2(1) = 1815$, p < .001).

This observation could be taken to indicate that clinicians' preferences for certain items influenced their item administration. This finding may however be caused by the fact that, in the current study design, individual interviewers do not necessarily meet the same group of patients, meaning that clinicians may have clinically relevant reasons for holding different item preferences. This possibility was accounted for in the next modeling step (Model 3), in which patient preliminary scores was included in the model as a cross-interaction with the item level variable. If clinicians follow the guidelines for the interview, individual patients are expected to be administered items from item levels adjacent to their assigned preliminary scores. As such, interviewers who administered the interview to patients who were assigned high/low preliminary scores, may naturally tend to administer more/fewer items from certain item levels, when adapting the interview to individual patients.

This interpretation is supported by findings of the next modeling step. When introducing 1) patient preliminary scores, 2) item level, and 3) the cross interaction between patient preliminary scores and item level to the model, the relative variance component of the item-by-interviewer interaction term dropped from 14% to 6%. This finding indicates that more than half of this variance could be accounted for by the introduction of these fixed-effects variables. The cross-interaction between patient preliminary scores and item level accounted for 34% of the variance in item administration ($\chi^2(3) = 7,020$, p < .001). Figure 2 provides some insight into the nature of this interaction. This figure shows the likelihood of items from levels 0, 1, 2, 3, and 4 being administered to patients assigned a preliminary score of 0 (represented by the dashed line) or 4 (represented by the full line) respectively. As can be seen from this figure, patients were much more likely to be administered items from a level close to their assigned preliminary score.

The addition of item covariates in the final modeling step (Model 4), led to a minor increase in the overall variance in

item administration explained by our model, as indicated by a 2% increase in $R^2$ from Model 3 to Model 4. Finding only a minor increase in $R^2$ was to be expected, as these latter, more incidental item characteristics ideally should have little to no influence on item administration, compared to expected patient severity. Surprisingly, however, the effects of these item characteristics on item administration were all significant, indicating that longer items of a level for which more alternative items are listed and items that appear early in a list of items, have a slightly higher chance of being administered when comparing items that are similar in all other aspects. Combining the random and fixed effects, this final model accounted for 67% of the variance in item administration.

## Discussion

This study examined clinicians' utilization of the funnel structure of the SCID-5-AMPD-I and found that clinicians substantially shortened the interview by administering on average only half of the 168 level-specific items available to them. Generally, items were selected from a subset of 2-3 adjacent levels under each subdomain, indicating that clinicians actively explored multiple levels which might be suitable to a given patient, but did not actively administer interview content from all levels under each subdomain. The administration of interview content was fairly tailored to individual patient cases, with no two interviews containing exactly the same set of administered items, and any two interviews sharing, on average, only half the administered content with each other. The selection of interview items was strongly impacted by clinicians' initial estimations of patient severity, such that clinicians tended to administer items corresponding to the preliminary scores assigned to the patients. The effect of more superficial item characteristics and aspects of the interview's internal organization was comparatively marginal. Together, these findings suggest: 1) That the funnel structure promotes efficient assessments (as the number of administered items is reduced by 50% and the number of levels assessed under each subdomain is reduced by 40%); 2) The funnel structure promotes individually tailored assessments (in the sense that individual interviews differ substantially in administered interview content); 3) Clinicians' decisions about which part of the interview to administer are guided by relevant patient characteristics (i.e., preliminary scores assigned to the patient at an early stage of the interview, representing expected patient severity).

Given that the SCID-5-AMPD-I has been found to have good to excellent interrater reliability, our findings suggest that non-fixed administration procedures such as the funnel structure of the SCID-5-AMPD-I can be implemented without necessarily representing a threat to reliability. These findings may pave the way for the development of innovative assessment procedures such as computerized adaptive tests (CATs). In fact, the non-fixed administration procedures endorsed in the SCID-5-AMPD-I are, in several respects, reminiscent of CATs. Item administration based in CATs is tailored to the patient on the fly, which is very

similar to the SCID-5-AMPD-I procedure. More specifically, a CAT selects items from an item bank based on information about the items as well as the patient, while continuously updating the estimation of the patient's score on the trait being measured (Chang, 2015; van der Linden & Glas, 2000). This means that individual patients are administered personalized item sets based on their answers to previously administered items. CATs are widely used in educational contexts and have, over the last few years, become increasingly popular in the field of clinical assessment (Gibbons et al., 2016; Gibbons et al., 2014; Paap et al., 2019; Smits et al., 2018). The clear advantage of a CAT, over linear assessments, is its ability to produce efficient and individually tailored assessments without compromising the reliability or comparability of scores across subjects (Braeken & Paap, 2020; Chang, 2015; Paap et al., 2019; van der Linden & Glas, 2000).

However, a very important difference between the SCID-5-AMPD-I and a CAT is the manner in which the estimation process is conducted. For a fixed-precision CAT, the item selection is guided by an algorithm using item and person characteristics estimated with statistical models and large datasets (Chang, 2015; Smits et al., 2018; van der Linden & Glas, 2000). When selecting the next item for administration, the algorithm typically selects the item which will produce the greatest reduction in measurement error for the estimated patient score. The test continues until the standard error of the estimated score falls below a predetermined threshold. For the SCID-5-AMPD-I on the other hand, the clinician is asked to perform a similar task based on their clinical expertise and relevant patient experience. It is therefore important to examine which factors impact clinicians' item selection, in order to verify that it is informed by relevant patient characteristics rather than other, unrelated factors. The manual for the SCID-5-AMPD-I specifies that clinicians should decide which levels to assess under each subdomain based on a) their initial evaluation of the patient's personality functioning, as indicated by the preliminary score assigned at the beginning of the interview; and b) patient answers to screening questions available at the start of each subdomain (First et al., 2018a). Clinicians are instructed to integrate this information with any observations they make about the patient and information gathered throughout the interview to decide which level of impairment is best suited to describe the patient.

Our findings suggest that clinicians in our sample followed these instructions quite well. At the same time, however, we found potential indications that other factors, unrelated to the patient, also influenced item selection. Out of the ten items with the most missing data on item administration, eight were items exploring narcissistic or antisocial traits and tendencies. Although this may have been a coincidence, previous studies have suggested that clinicians may avoid administering confrontational items when provided with the freedom to select items in the SCID-5-AMPD-I. Heltne et al. (2022), for instance, reported that clinicians in the NorAMP sample described skipping items they considered overly complex or confrontational. Such preferential

item selection has also been suggested to have affected the interrater reliability of the instrument in the original NorAMP study. Here, the subdomain 'Mutual regard reflected in behavior' had the lowest reliability with an ICC of .24 (Buer Christensen et al., 2018). These authors suggested one possible reason for this finding may be that raters omitted confrontational or morally laden questions listed under level 2 of this subdomain, and prematurely skipped from level 1 to level 3. In our own analysis, we found that the subdomain of 'Mutual regard reflected in behavior' had by far the greatest number of non-adjacent level administration with 26 instances, 25 of which represented the clinician administering items from level 1 and level 3 without administering any items from level 2. Though these findings may be the result of culturally based tendencies specific to the Norwegian samples used in these studies, these findings highlight the importance of examining factors that impact item selection for the SCID-5-AMPD-I as well as for any other interview which adopts similar, non-fixed administration procedures.

Our study also found that superficial item characteristics related to the content and structure of the interview may impact item selection. Although the effects were relatively minor, clinicians were more likely to administer longer items and items from level-subdomain combination with multiple available items. Clinicians were also more likely to administer items near the top of a list of available items than those listed further down. In the current version of the SCID-5-AMPD-I, the number of items listed under the levels of each subdomain range from 1 to 6. At the time of writing, the developers of the interview have not offered formal explanation or reasoning for this variability in available items or the ordering of items either on the official website of the instrument (Columbia University, 2020) or in the manual for the instrument (First et al., 2018a). Minor adjustments to the interview's content and structure could prevent these more superficial elements from unduly affecting item selection. For future editions of the SCID-5-AMPD-I, we therefore suggest that developers present a fixed number of items for each level, under each subdomain, and that these items are ordered by their relative usefulness for evaluating the level description they are meant to assess. To realize this aim, we recommend surveying experienced clinicians and asking them to order items according to their relative suitability to assess the core aspects of corresponding level descriptions.

Although the impact of preferential item selection and structural elements of the interview were small in the current sample, it is difficult to know how well these findings will generalize to a naturalistic clinical setting. Most clinicians involved in the NorAMP had previous experience with PD patients. They also received a two-day training seminar which covered relevant theory and practice exercises. Furthermore, these clinicians attended regular calibration meetings together. Several of these clinicians have cited training and calibration meetings as being essential for their ability to administer the interview correctly (Heltne et al., 2022). In a naturalistic clinical setting, however, access to

training and level of experience of clinicians will vary. Untrained novice clinicians may have trouble setting an accurate preliminary score to guide further item selection, leading them to have to administer more levels under each subdomain before arriving at an appropriate score for their patients. Replicating the findings of the current study in more naturalistic clinical settings is therefore needed in order to determine how important training is to clinicians' ability to benefit from the SCID-5-AMPD-I funnel structure.

Even if similar results can be obtained in samples of untrained raters, it is important to acknowledge that the funnel structure of the SCID-5-AMPD-I may introduce an additional risk of confirmation bias as compared to standard fixed administration procedures in which all items are administered to all patients. When an initial impression about the patient is not only allowed to, but rather is supposed to, affect decisions about which sections of the interview to administer, there is a risk that clinicians will be more likely to selectively seek information which confirms rather than rejects this initial impression. While there are elements in the interview's administration guidelines which could limit the impact of administration bias (e.g., the instruction to assess increasing levels until the level under assessment no longer applies), it was beyond the scope of the current study to evaluate the impact of confirmation bias on the administration and scoring of the SCID-5-AMPD-I. Given that the risk of confirmation bias in clinical decision making and diagnostic assessments has been documented in numerous studies (e.g., Crumlish & Kelly, 2009; Mendel et al., 2011; Strohmer & Shivy, 1994) and may be associated with erroneous diagnostic conclusions (Mendel et al., 2011), it is important to address this issue in future research. Before fully embracing the funnel structure of the SCID-5-AMPD-I, we therefore strongly recommend evaluating whether the instrument introduces confirmation bias.

One way to go about it would be to compare total scores between repeated interviews where patients are first administered the interview according to the funnel structure, and then the entire interview. Confirmation bias would then be expected to lead to low predictive validity between these two assessments. Alternatively, one could compare the sensitivity and specificity between funnel-structure and full-administration interviews. It is important that such investigations be carried out before choosing non-fixed administration procedures over more traditional linear procedures. Not doing so would be to ignore a potentially crucial limitation of this novel and innovative form of conducting semi-structured clinical interviews.

## Conclusion

Our results show that the funnel structure of the SCID-5-AMPD-I has the potential to enable clinicians to conduct more efficient and individually tailored assessments than is possible with fixed administration procedures, in which all patients are administered all interview items. The benefits of these non-fixed administration procedures of the SCID-5-AMPD-I does not appear to negatively impact the instrument's reliability. Similar non-fixed administration procedures may thus represent a promising avenue for improving the efficiency and individualization of other semi-structured interviews. When similar procedures for other instruments are adopted, clear administration guidelines and adequate training are recommended in order to achieve optimal results. Furthermore, it is important to verify that item administration is informed by relevant factors. The current study can serve as a template for conducting such verifications. Although beyond the scope of the current study, it is important to address the risk of confirmation bias which may arise, when clinicians are instructed to use their initial impressions of the patient's condition to guide content selection in a semi-structured interview. We have therefore provided specific suggestions on how this matter can be explored in future studies.

## Declaration of interest

The authors declare there are no conflicts of interest.

## Funding

## ORCID

Aleksander Heltne http://orcid.org/0000-0002-3524-203X
Johan Braeken http://orcid.org/0000-0002-2119-3222
Benjamin Hummelen http://orcid.org/0000-0002-8717-8076
Sara Germans Selvik http://orcid.org/0000-0001-9921-8971
Tore Buer Christensen http://orcid.org/0000-0002-9908-8182
Muirne C. S. Paap http://orcid.org/0000-0002-1173-7070

## Data availability statement

The data that support the findings of this study are available from the third author, Benjamin Hummelen, upon reasonable request, after approval by the Data Protection Manager at Oslo University Hospital.

## References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders: DSM-IV*. American Psychiatric Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed.). American Psychiatric Association. https://doi.org/10.1176/appi.books.9780890425596

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1– 48. https://doi.org/10.18637/jss.v067.i01

Braeken, J., & Paap, M. C. S. (2020). Making fixed-precision between-item multidimensional computerized adaptive tests even shorter by reducing the asymmetry between selection and stopping rules. *Applied Psychological Measurement*, *44*(7–8), 531–547. https://doi.org/10.1177/0146621620932666

Buer Christensen, T., Hummelen, B., Paap, M. C. S., Eikenaes, I., Germans Selvik, S., Kvarstein, E., Pedersen, G., Bender, D. S., Skodol, A. E., & Nysaeter, T. E. (2019). Evaluation of diagnostic thresholds for criterion A in the alternative DSM-5 model for

personality disorders. *Journal of Personality Disorders*, 1–22. https://doi.org/10.1521/pedi_2019_33_455

Buer Christensen, T., Paap, M. C. S., Arnesen, M., Koritzinsky, K., Nysaeter, T. E., Eikenaes, I., Germans Selvik, S., Walther, K., Torgersen, S., Bender, D. S., Skodol, A. E., Kvarstein, E., Pedersen, G., & Hummelen, B. (2018). Interrater reliability of the structured clinical interview for the DSM-5 alternative model of personality disorders module i: Level of personality functioning scale. *Journal of Personality Assessment*, 100(6), 630–641. https://doi.org/10.1080/00223891.2018.1483377

Chang, H. H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, 80(1), 1–20. https://doi.org/10.1007/s11336-014-9401-5

Clark, L. A. (2007). Assessment and diagnosis of personality disorder: Perennial issues and an emerging reconceptualization. *Annual Review of Psychology*, 58, 227–257. https://doi.org/10.1146/annurev.psych.57.102904.190200

Columbia University. (2020). *Structured clinical interview for DSM Disorders (SCID) FAQ*. Columbia University Department of Psychiatry. Retrieved July 15, 2020, from https://www.columbiapsychiatry.org/research/research-labs/diagnostic-and-assessment-lab/structured-clinical-interview-dsm-disorders-12

Crumlish, N., & Kelly, B. D. (2009). How psychiatrists think. *Advances in Psychiatric Treatment*, 15(1), 72–79. https://doi.org/10.1192/apt.bp.107.005298

Ekselius, L., Lindström, E., von Knorring, L., Bodlund, O., & Kullgren, G. (1994). SCID II interviews and the SCID screen questionnaire as diagnostic tools for personality disorders in DSM-III-R. *Acta Psychiatrica Scandinavica*, 90(2), 120–123. https://doi.org/10.1111/j.1600-0447.1994.tb01566.x

First, M. B., & Gibbon, (2004). The structured clinical interview for DSM-IV axis I disorders (SCID-I) and the structured clinical interview for DSM-IV axis II disorders (SCID-II). In *Comprehensive handbook of psychological assessment, Vol. 2: Personality assessment* (pp. 134–143). John Wiley & Sons Inc.

First, M. B., Skodol, A. E., Bender, D. S., & Oldham, J. M. (2018a). *User's guide for the SCID-5-AMPD: Structured clinical interview for the DSM-5 alternative model for personality disorders*. American Psychiatric Association Publishing.

First, M. B., Skodol, A. E., Bender, D. S., & Oldham, J. M. (2018b). *Structured clinical interview for the DSM-5 alternative model for personality disorders module I: Structured clinical interview for the level of personality functioning scale*. American Psychiatric Association.

First, M. B., Williams, J. B. W., Benjamin, L. S., & Spitzer, R. L. (1997). *SCID-II: Structured clinical interview for DSM-IV AXis II personality disorders: User's guide*. American Psychiatric Press.

First, M. B., Williams, J. B. W., Benjamin, L. S., & Spitzer, R. L. (2015a). *Structured clinical interview for DSM-5 personality disorder*. American Psychiatric Association.

First, M. B., Williams, J. B. W., Benjamin, L. S., & Spitzer, R. L. (2015b). *User's guide for the SCID-5-PD (structured clinical interview for DSM-5 personality disorder)*. American Psychiatric Association.

Fossati, A., & Somma, A. (2021). The assessment of personality pathology in adolescence from the perspective of the Alternative DSM-5 model for personality disorder. *Current Opinion in Psychology*, 37, 39–43. https://doi.org/10.1016/j.copsyc.2020.07.015

Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized adaptive diagnosis and testing of mental health disorders. *Annual Review of Clinical Psychology*, 12, 83–104. https://doi.org/10.1146/annurev-clinpsy-021815-093634

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2014). Development of the CAT-ANX: A computerized adaptive test for anxiety. *The American Journal of Psychiatry*, 171(2), 187–194. https://doi.org/10.1176/appi.ajp.2013.13020178

Heltne, A., Bode, C., Hummelen, B., Falkum, E., Germans Selvik, S., & Paap, M. C. S. (2022). Norwegian clinicians' experiences of learnability and usability of SCID-II, SCID-5-PD and SCID-5-AMPD-I interviews: A sequential multi-group qualitative approach. *Journal of Personality Assessment*, 104(5), 599–612. https://doi.org/10.1080/00223891.2021.1975726

Hummelen, B., Braeken, J., Buer Christensen, T., Nysaeter, T. E., Germans Selvik, S., Walther, K., Pedersen, G., Eikenaes, I., & Paap, M. C. S. (2021). A psychometric analysis of the structured clinical interview for the DSM-5 alternative model for personality disorders module I (SCID-5-AMPD-I): Level of personality functioning scale. *Assessment*, 28(5), 1320–1333. https://doi.org/10.1177/1073191120967972

Kampe, L., Zimmermann, J., Bender, D., Caligor, E., Borowski, A. L., Ehrenthal, J. C., Benecke, C., & Hörz-Sagstetter, S. (2018). Comparison of the structured DSM-5 clinical interview for the level of personality functioning scale with the structured interview of personality organization. *Journal of Personality Assessment*, 100(6), 642–649. https://doi.org/10.1080/00223891.2018.1489257

Meisner, M. W., Bach, B., Lenzenweger, M. F., Møller, L., Haahr, U. H., Petersen, L. S., Kongerslev, M. T., & Simonsen, E. (2022). Reconceptualization of borderline conditions through the lens of the alternative model of personality disorders. *Personality Disorders: Theory, Research, and Treatment*, 13(3), 266–276. https://doi.org/10.1037/per0000502

Mendel, R., Traut-Mattausch, E., Jonas, E., Leucht, S., Kane, J. M., Maino, K., Kissling, W., & Hamann, J. (2011). Confirmation bias: Why psychiatrists stick to wrong preliminary diagnoses. *Psychological Medicine*, 41(12), 2651–2659. https://doi.org/10.1017/S0033291711000808

Miller, P. R., Dasher, R., Collins, R., Griffiths, P., & Brown, F. (2001). Inpatient diagnostic assessments: 1. Accuracy of structured vs. unstructured interviews. *Psychiatry Research*, 105(3), 255–264. https://doi.org/10.1016/S0165-1781(01)00317-1

Mooney, C. Z. (1997). *Monte Carlo simulation*. SAGE Publications, Inc. https://doi.org/10.4135/9781412985116

Mueller, A. E., & Segal, D. L. (2015). Structured versus Semistructured versus Unstructured Interviews. In *The Encyclopedia of Clinical Psychology* (pp. 1–7). https://doi.org/10.1002/9781118625392.wbecp069

Ottosson, H., Bodlund, O., Ekselius, L., Grann, M., von Knorring, L., Kullgren, G., Lindström, E., & Söderberg, S. (1998). DSM-IV and ICD-10 personality disorders: A comparison of a self-report questionnaire (DIP-Q) with a structured interview. *European Psychiatry*, 13(5), 246–253. https://doi.org/10.1016/S0924-9338(98)80013-8

Paap, M. C. S., Born, S., & Braeken, J. (2019). Measurement efficiency for fixed-precision multidimensional computerized adaptive tests: Comparing health measurement and educational testing using example banks. *Applied Psychological Measurement*, 43(1), 68–83. https://doi.org/10.1177/0146621618765719

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statisticsl Computing. https://www.R-project.org/

Saigh, P. A. (1992). Structured clinical interviews and the inferential process. *Journal of School Psychology*, 30(2), 141–149. https://doi.org/10.1016/0022-4405(92)90026-2

Smits, N., Paap, M. C. S., & Böhnke, J. R. (2018). Some recommendations for developing multidimensional computerized adaptive tests for patient-reported outcomes. *Quality of Life Research*, 27(4), 1055–1063. https://doi.org/10.1007/s11136-018-1821-8

Strohmer, D. C., & Shivy, V. A. (1994). Bias in counselor hypothesis testing: Testing the robustness of counselor confirmatory bias. *Journal of Counseling & Development*, 73(2), 191–197. https://doi.org/10.1002/j.1556-6676.1994.tb01735.x

Stuart, A. L., Pasco, J. A., Jacka, F. N., Brennan, S. L., Berk, M., & Williams, L. J. (2014). Comparison of self-report and structured clinical interview in the identification of depression. *Comprehensive Psychiatry*, 55(4), 866–869. https://doi.org/10.1016/j.comppsych.2013.12.019

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Kluwer Academic Publishers.

Widiger, T. A., & Samuel, D. B. (2005). Evidence-based assessment of personality disorders. *Psychological Assessment*, 17(3), 278–287. https://doi.org/10.1037/1040-3590.17.3.278